Article

Credible impact evaluation in complex contexts: confirmatory and exploratory approaches.

James Copestake

(Centre for Development Studies) University of Bath

Corresponding author:

James Copestake, Centre for Development Studies,

University of Bath, Bath, BA2 7AY, UK.

Email: j.g.copestake@bath.ac.uk

[James Copestake is Professor of International Development at the University of Bath, UK. His research interests encompass: rural development and agrarian change; development finance, microfinance and aid management; the definition and measurement of well-being, development and poverty; the political economy of development.]

Abstract

Debate continues over how best international development agencies can evaluate the impact of actions intended to reduce poverty, insecurity and vulnerability in diverse and complex contexts. There are strong ethical grounds for simply asking those intended to benefit what happened to them, but it is not obvious how to do so in a way that is sufficiently free from bias in favour of confirming what is expected. This article considers scope for addressing this problem by minimising the prior knowledge participants have of what is being evaluated. The tensions between more confirmatory and exploratory methodological approaches are reviewed in the light of experience of designing and piloting a qualitative impact assessment protocol for evaluating NGO interventions in complex rural livelihood transformations. The paper concludes that resolving these tensions entails using mixed methodologies, and that the importance attached to exploratory (nested within confirmatory) approaches depends on contextual complexity, the type of evidence sought and the level of trust between stakeholders.

Key words

Impact evaluation, qualitative methods, development practice, complexity, confirmation bias, pro-project bias

Évaluation d'impact crédible en contextes complexes: approches confirmatoires et exploratoires.

Cet article examine la controverse concernant la manière dont les agences de développement international évaluent l'impact des actions visant à réduire la pauvreté, l'insécurité et la vulnérabilité dans des contextes divers et complexes. Il existe des bases éthiques selon lesquelles il suffirait de demander aux bénéficiaires quels sont les changements occasionnés par l'intervention, mais il reste encore à voir comment ceci pourrait ne pas être biaisé en faveur de confirmer la réponse attendue. Cet article examine si minimiser la connaissance préalable des participants sur ce qui est évalué pourrait réduire l'ampleur de ce problème. L'auteur met en évidence les tensions entre les approches méthodologiques confirmatoires et exploratoires dans le cadre d'expériences de mise en œuvre de protocoles d'évaluation d'impact qualitative pour évaluer les interventions d'ONGs en milieu rural complexe.

Introduction

This paper addresses the perennial question of how international development agencies, who make major investments across vast distances and areas, can credibly evaluate the impact of their diverse policies, programmes, strategies, projects and practices relative to their stated development goals. The paper draws particularly on collaborative action research with two NGOs, Self Help Africa (SHA) and Farm Africa, who confront the problem of how best to evaluate annual spending of around £20 million across twelve countries relative to their aim to support sustainable agricultural growth and improved food security.

These and other development activities confront political and methodological issues familiar to those interested in impact evaluation of public policies and programmes in many other contexts. These include the contribution of impact evaluation both to organisational learning and building legitimacy through improved upward and downward transparency and accountability. These are linked to wider debates in international development over the role of results-based management culture and the politics of evidence (e.g. Gulrajani, 2010; Eyben, 2013). In contrast, this paper focuses on narrower methodological issues, particularly the problem of attribution, or how particular outcomes - such as improvements in household level food security - can reliably be linked to specific projects, interventions or mechanisms in different contexts. But technical and political aspects of impact evaluation can never be fully separated from each other, and the methodological issues discussed in this paper are framed by a concern with upward accountability on the part of non-government development agencies reliant on funding from government and the public in more affluent countries. This influences, for example,

what is meant by ‘credible’ impact evaluation and how much money it is realistic to be able to spend on it.

Before turning to the question of attribution, a brief reference is appropriate to the principle of directly involving intended primary beneficiaries of a development intervention in its evaluation. If we are interested in finding out whether particular men, women or children are less hungry as a result of some action it seems common-sense just to ask them. But even putting aside problems of construct validity (over the definition of hunger, for example) it is not obvious how easily they will be able to attribute changes in their experience to specific activities. And there may also be reasons to doubt the reliability of their responses, including *confirmation bias* (Haidt, 2012:93) or a tendency to “anchor” their response to what is familiar or expected (Kahneman, 2011). In this paper I will also use the term *pro-project bias* to refer to the possibility that someone consciously or otherwise conceals or distorts what they think they know about an activity in the hope that doing so will reinforce the case for keeping it going. The instrumental value of asking people directly about attribution (benefitting from their close knowledge, but subject to such bias) is ultimately an empirical question, albeit a hard one. However, there are also ethical reasons for asking people directly, captured in the rallying cry of the international disability movement ‘nothing about us without us’, and explored in the context of international aid by Anderson *et al.* (2012). While *ex post* consultation may be a weak substitute for intended beneficiaries having a role in planning and sanctioning work done in their name, neither is lack of such involvement a reason for excluding them from evaluation. In sum, direct involvement of at least some intended beneficiaries in impact evaluation is ethically

important, even if it presents methodological challenges. How then, (if at all) can self-reported assessment of impact and its causes assist directly in tackling the attribution problem that lies at the heart of impact evaluation?

Approaches to impact evaluation

In relating the issue of self-reported impact assessment to the wider literature I will make two heuristic distinctions: between more quantitative and qualitative approaches to evaluating impact, and between more exploratory and confirmatory approaches. There is of course no shortage of literature on the first of these, including recent reviews of the status of qualitative and mixed approaches in an international development context (Shaffer, 2013; Stern *et al*. 2012; White and Phillips, 2012). The dominant quantitative approach can be linked to an axiomatic view of impact as the difference in the value of an outcome indicator (Y_1) for a given population after a particular intervention or 'treatment' (X) compared to what the value would have been for the same population if the treatment had not occurred (Y_0) (White, 2010:154). Putting aside the problem of consistent measurement of X and Y, a central issue is then how to establish a plausible counterfactual. If the evaluator can make a large number of observations of X and Y then they can draw on well-known quantitative approaches to address this problem, including the use of randomized control designs. In contrast, this paper focuses on the scope for more qualitative and 'small n' approaches, primarily on the grounds that there has been less clarity, consistency and consensus about how best to employ these in evaluating the impact of development activities (Stern *et al.*, 2012:1; White and Phillips, 2012:5). In passing, it is germane to the argument of this paper to note that even

randomization is not of itself a guarantee against pro-project bias in estimates of impact (White, 2010:156). This is particularly the case if Y is obtained from respondents (and/or by researchers) who are not blind to whether they belong to the treatment or control sample, and may therefore be prone to different degrees of response bias, including Hawthorne and John Henry effects (Duvendack *et al.* 2011).

Among various criticisms of quantitative approaches that rely on experimental or quasi-experimental designs (e.g. Deaton, 2010; Cartwright, 2011; Picciotto, 2012) perhaps the most important concern the practical feasibility of addressing the practical threats to internal validity identified.[1] In an immensely complex, diverse, fast changing, emergent and recursive social world many researchers have argued that it is simply too slow and expensive to generate sufficient data in this way to be that useful. It may be possible to measure a large vector of variables **Y** for a given population and time period, and to demonstrate how they are affected by exposure to a vector of interventions or treatments **X**. But each set of results is specific in time and space to a vector of confounding or contextual variables (**Z**) that is too small or too quickly becomes outdated in history (see for example, Pawson and Tilley, 1994, and the ensuing exchange with Bennett). Realist evaluation can be viewed specifically as a counterpoint to this: emphasising the need for a cumulative process of broadening understanding of context-mechanism-outcome interactions or knowledge of "…what works for whom in what circumstances, in what respects, over which duration… and why" (Pawson and Manzano-Santaella, 2012:177). Its pursuit of realism can be viewed as being achieved at the expense of the precision gained from artificially restricting variation

in treatment and contextual vectors in order to generate statistically significant results (Levins, 1966).

Realist evaluation constitutes only one possible counterpoint to positivist approaches, and can be linked to a range of 'theory driven' and 'theory of change' approaches in emphasising the importance of building evaluation on prior elucidation of programme theory linking mechanism to outcomes (Pawson and Manzano-Santaella, 2012:178; Mayne, 2012; Ton, 2012). White and Phillips (2012:4&34) identify a cluster of qualitative approaches sharing a common core that "involves the specification of a theory of change together with a number of further alternative causal hypotheses." They also identify a second cluster of approaches that "place stakeholder participation at the heart of data collection and analysis" but conclude that these do not make causal explanation their primary goal, are prone to various biases arising from their reliance on stakeholder perception, and are most usefully employed only as one element within a wider evaluation framework (p.21).[2]

This brings me to a second and less widely discussed heuristic distinction this is the main subject of this paper: between confirmatory and exploratory approaches to impact evaluation. At first glance this distinction seems close to the one between the two groups drawn by White and Phillips. Their first group can be viewed as confirmatory in the sense that it seeks evidence to either validate or challenge the researcher's prior theory, and thereby downplays those of other stakeholders. If so, then this is contrasted with participatory approaches which *a priori* give more weight to the perceptions of other stakeholders. An alternative classification would be to contrast *confirmatory* small *n* approaches with those that are more open-ended and *exploratory* in the sense of explicitly limiting prior theorisation on the part of the

researcher. Combining the two distinctions leaves us with a set of four contrasting approaches as depicted in Table 1.

Table 1: A typology of approaches to 'small n' impact evaluation

	Privilege the professional judgement of the evaluator	Give more weight to the views of other participants or project stakeholders
More confirmatory	I	II
More exploratory	III	IV

White and Philip's assertion that participatory approaches II and IV are more prone to bias than I and III seems to boil down to an axiomatic statement of faith in the professional judgement of 'independent' evaluators, including their capacity to rise above the knowledge claims of other stakeholders. This raises questions about the wider politics of impact evaluation that fall beyond the scope of this paper. However, even without challenging the status of the evaluator (and thereby ignoring the second column) there remains the question of how - and in what contexts - the scope of impact evaluation methods should be extended beyond testing predetermined hypotheses to accommodate more open-ended, naturalistic and inductive methods, allowing theory to emerge through interaction between prior theory and elicited stakeholder perceptions. How, how far, and in what context is it feasible and appropriate to incorporate more exploratory qualitative and mixed social research methods and standards to the more tightly scripted world of development evaluation?[3]

To illustrate the scope for more exploratory approaches it is useful to review an already highly developed method that falls in category IV in the typology of Table

1.[4] Participatory Assessment of Development (PADev) has been developed over six years by a consortium of Dutch and West African NGOs and researchers coordinated by Ton Dietz at the African Studies Centre at the University of Leiden (Dietz and PADev Team, 2013). A core objective of PADev, as set out in published guidelines that build on eleven field workshops (PADev, 2012), is to address the limitations of impact evaluations that artificially focus on the activities of just one project or agency in areas that have been affected by multiple and overlapping interventions. This problem is particularly acute for those who regard development as the outcome of highly interconnected, hierarchically nested and multi-dimensional systems of activity, and are sceptical of how far it is possible or meaningful to identify (still less quantify) the isolated causal contribution of just one intervention (Schiefer, 2008; Bevan, 2013). PADev's solution to this problem is collective action in the form of joint evaluation of all development interventions in a specified locality over a long period of time. The guidelines offer a suite of interconnected and structured participatory exercises aimed at enabling a reasonably representative cross-section of the population to construct nothing less than a joint history of development in and of the locality over a period of up to thirty years. These activities include constructing a time line of major events, wealth ranking, evaluating long-term changes in wellbeing, identifying significant interventions and ranking them according to both immediate and long-term impact. While open-ended and exploratory, the use of predesigned Excel tables for recording outputs aids relatively rapid and structured generation of findings for evaluation purposes.

The PADev approach inevitably entails expert facilitation, but also goes to considerable lengths to reflect the views of a cross-section of people, including

intended and unintended beneficiaries, rich and poor. At the same time, by moving the spotlight away from any one agency it can claim to go some way towards reducing pro-project bias, even if the possibility of some positive bias towards externally sponsored development activities in general is hard to completely eliminate, alongside deference to the outsiders perceived to be mediating the assessment (Dietz, 2012:233-4). A third challenge facing this approach is that it entails producing a public good and hence is susceptible to free rider problems. More specifically, individual development agencies may struggle to justify investing in an impact evaluation that generates multiple findings, only a small component of which relate specifically to their own activities: what is gained in terms of contextualisation and balance having to be offset against less detail and precision about their own activities. Overcoming this problem by pooling the evaluation resources of several agencies operating in particular places is possible, as the case of joint donor funding of the WIDE project in Ethiopia also illustrates (Bevan, 2013). But it is difficult, and it is for this reason that this paper advocates research into a broader range of more exploratory impact evaluation methods that retain an ultimate focus on the impact of one particular agency or project. Before doing so, it is worth briefly clarifying what is meant by credible impact evaluation, and discussing in more depth possible criteria for assessing the strengths and weaknesses of different approaches.

Defining credible impact evaluation

White (2010:154) notes that the term impact evaluation is widely used to refer both to any discussion of outcome and impact indicators, and more narrowly to studies

that explicitly seek to attribute outcomes to a specified intervention. This paper adopts the second definition, while at the same time taking a broad view about what constitutes a sufficient level of scientific rigour in addressing the attribution problem. Doing so allows for the possibility that specific *methods* (including those within a positivist tradition) can be nested within broader (including interpretive) *approaches*. This acknowledges that attributing impact is only one question (if perhaps the most important) that impact evaluations need to answer – others including how an intervention works, and whether it constitutes value for money (Stern et *al.*, 2012:36).

As a servant of action in a constantly changing historical context the scientific rigour of impact evaluation also has to be weighed alongside cost, timeliness and fitness to purpose. Without rejecting the quest for consensus about what constitutes quality in qualitative research (e.g. as discussed Hammersley, 2013:83) this leads me to a preference for the term credibility over scientific rigour, echoing the broader distinction between reasonableness and rationality (McGilchrist, 2010).[5] By credibility, I refer to one party being able to offer a sufficient combination of evidence and explanation to convince another party that a proposition is reasonable in the sense of being sufficiently plausible to act upon – not rational in a logical sense, perhaps, but neither irrational. While this emphasises the importance of context and trust, the rigour with which conclusions about impact are logically derived from stated evidence and assumptions is also clearly important. A common distinction here is between the validity of an approach, and the reliability of results arising from its application in a particular context (e.g. Lewis and Ritchie, 2003:270). Given that no IE can ever be replicated in precisely the same context or setting in

time and place, the value of this distinction to qualitative research is questionable, and in the context of international development it is rather analogous to nebulous (if not disingenuous) attempts to distinguish policy from implementation, or theory from practice. The underlying problem here is the complexity of the context. By this I mean that the influence of X on Y is confounded by factors **Z** that are impossible fully to enumerate, of uncertain or highly variable value, difficult to separate, and/or impossible fully to control. Additional complexity arises if the nature and value of X and/or Y is also uncertain.

An alternative approach to defining credibility with respect to impact evaluation is to specify what constitutes reasonable evidence of causation. For example, an evaluator's claim to establishing impact (i.e. X causing Y in particular contexts) might be regarded as being credible if: (a) there is strong evidence that X and Y happened in such contexts, (b) X is described by a diverse range of stakeholders as having been a necessary cause of Y in those contexts[6], (c) their explanations of the mechanism by which X caused Y in those contexts are independently arrived at and mutually consistent, (d) the counter-hypothesis that they have other reasons for making the statement can reasonably be refuted. The point is not to secure universal agreement, but to be as clear and precise as possible about what can reasonably be expected in a given context. For example, our emphasis here being on qualitative methods, the definition excludes the requirement for (e) evidence of how *much* Y varies according to exposure to X. The idea of credible causation, based on reasonableness, can be further elaborated by specifying minimum conditions for mitigating the risks of systematic bias. The definition above, for example, proposes structures and processes of evaluation that

reduce the plausibility of complicity among different stakeholders. This falls short of scientific certainty, but in complex contexts it is often as much as we can hope for, particularly given the possibility that efforts to aim higher may be counterproductive in terms of cost, timeliness and policy relevance. In other words, I am not suggesting that this definition is universal or even widely accepted, rather that it is a realistic one in contexts where overcoming the attribution problem is particularly difficult.

Case study: the ART Project

Description and rationale

This section explores the core issue of the tension between confirmatory and exploratory aspects of impact evaluation by reviewing the design and piloting of a qualitative impact protocol (QUIP) intended to deliver 'good enough' evidence of the impact of NGO activities (hereafter referred to as projects) to strengthen smallholder agricultural livelihoods in the context of the complex rural transformations taking place in many parts of Africa.[7]

Initial testing was conducted with two projects in Malawi being implemented by the NGO Self-Help Africa with financial support from Irish Aid and UK Aid. These both comprised a battery of interventions (including new seed varieties, training in conservation farming, beekeeping, livestock promotion, small scale irrigation and microfinance) intended to strengthen household level food security and wellbeing. The context of the impact assessment task is complex in the sense – defined earlier in this paper - of the presence of numerous, interconnected, uncertain and hard-to-

measure confounding variables **Z** (agronomic, climatic and commercial) affecting the casual links between multi-faceted NGO interventions **X** and intended impacts **Y**.

In contrast to quantitative impact assessment methods, the QUIP sets out to generate differentiated empirical evidence of impact based on narrative causal statements of intended project beneficiaries without the requirement to interview a control group. Evidence of attribution is sought through respondents' own account of causal mechanisms linking **X** to **Y** alongside **Z** rather than by relying on statistical inference based on variable exposure to **X**. More specifically, the research proposal hypothesised scope for improvement in the credibility of such impact assessment through close attention to methodological details, including sample selection, framing of interviews, structuring of questions within data collection instruments, triangulation, data analysis and procedural transparency. Draft guidelines for the QUIP were presented to a methodology workshop held in June 2013 and attended by staff from the University of Bath, the University of Malawi, Self Help Africa, Farm Africa, Evidence for Development, Oxfam UK and Irish Aid.[8] Each section was subject to detailed discussion at the workshop, and further refined before and after pilot field testing for two SHA projects in Malawi in November 2013. The guidelines cover commissioning of impact assessment, its relationship to other impact evaluation activities, sample selection, data collection methods, briefing and debriefing the field researchers, facilitating interviews, data analysis, quality assurance and use of findings.[9]

Each pilot comprised eight households level interviews plus four focus group discussions conducted by two independent researchers over five days. Narrative data was recorded in the field on a paper pro-forma and then copied into an Excel

spreadsheet with an identical layout. The field researchers were contracted by the University of Bath (acting as lead evaluator) without any contact with the NGO or staff of the project being evaluated. Moreover, interviews and focus group discussions were set up without the researchers having any prior knowledge of the selected NGO project or the theory of change underpinning it - the idea of this procedural blinding being to enhance credibility of findings by greatly reducing the risk of pro-project bias (see below). Instead the field team introduced themselves as conducting research into general changes in the rural livelihoods and food security of farmers in the selected area. Data thereby collected was then passed to staff at the University of Bath whose role was to identify cause-and-effect statements contained within it and to code them according to whether they (a) *explicitly* attributed impact to project activities, (b) made statements that were *implicitly* consistent with the project's theory of change, (c) referred to drivers of change that were *incidental* to project activities. Table 2 presents summary data from the two projects to illustrate how these statements were classified according to thematic domains (corresponding to those used to structure data collection) and whether respondents classified them as having a positive or negative effect on their well-being. Discussion of each domain concluded with additional closed questions (not shown). The analysis also explored in far more detail than shown here the specific causal mechanism cited by respondents. This analysis was then fed back to the NGO in the form of a ten page report along with an annex setting out the coded cause-and-effect statements in full. The University of Bath also debriefed the field team, confirming in this case that they were unsure of the identity of the two specific projects being evaluated even after completion of the data collection work.

Table 2: Frequency counts of causal statements obtained from QUIP pilots in Malawi.

Type of statement	Explicit		Implicit		Incidental		Unattributed	
Domain	+	-	+	-	+	-	+	-
Project 1								
Food production	5,2	0,0	2,1	0,3	1,0	1,2	0,0	0,0
Cash income	4,4	0,0	4,0	1,2	2,0	2,0	0,0	0,0
Cash spending	4,4	0,0	1,0	0,1	1,0	2,2	0,0	0,0
Food consumption	3,1	0,0	1,0	0,2	1,0	1,0	0,0	0,0
Relationships	4,1	0,0	1,0	0,0	2,0	0,2	0,0	0,0
Asset accumulation	2,2	0,0	0,1	1,1	2,0	0,0	0,0	0,0
Project 2								
Food production	1,0	0,0	2,0	1,0	0,0	4,4	0,0	0,0
Cash income	1,1	0,0	5,3	1,0	0,0	3,4	0,0	0,0
Cash spending	0,0	0,0	2,1	0,0	0,0	5,3	0,0	0,0
Food consumption	0,0	0,0	4,0	0,0	1,0	2,4	0,0	0,0
Relationships	0,0	0,0	0,3	0,0	3,3	1,2	0,0	0,0
Asset accumulation	0,0	0,0	3,0	0,0	0,2	0,3	0,0	0,0

Source: ART Project, primary data.

Notes: The first number in each cell refers to how many household interviews (out of eight per project) yielded such statements, and the second to how many focus groups did so (out of a maximum of four per project). The four 'types of statement' were defined as: 'explicit' = change explicitly attributed to the project or explicitly named project activities; 'implicit' = change confirming or refuting the specific mechanism or theory of change by which the project aims to achieve impact, but with no explicit reference to the project or named project activities; 'incidental' = change attributed to other forces, not related to activities included in the project's theory of change; 'unattributed' = change not attributed to any specific cause. Domains refer to sections of the interview and focus group schedules. Analysts classified statements as positive or negative according to the impact on respondents' wellbeing as expressed by respondents themselves; an option to classify responses as 'neutral' or unclear in its impact on the stated domain was also available, but was not utilised.

QUIP design issues

This section critically reflects on four sets of issues that emerged in the process of designing and piloting the QUIP. These concern: (a) research roles and relationships; (b) linking impact assessment and project monitoring; (c) timing of interviews and sample selection; (d) interview questions and data analysis.

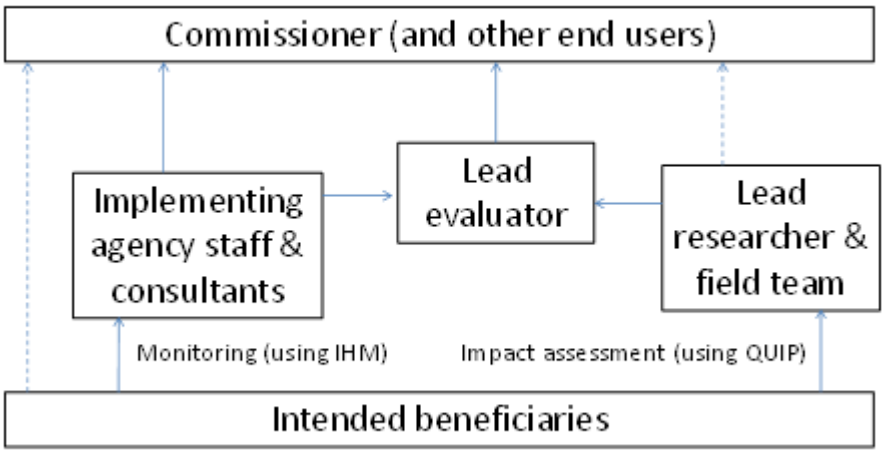
Research roles and relationships. A key idea behind the design of the QUIP was the goal of eliciting open-ended narrative about changes in the lives and livelihoods of respondents through which statements about the impact of a specific project are

volunteered without prompting thereby directly addressing the pro-project bias risk associated with self-reported attribution. With this in mind the protocol was designed to minimise contact between field researchers contracted to collect the QUIP data and project staff, so that the former are as far as possible unaware of the nature of the project being evaluated and the theory of change underpinning it. This emphasis on avoiding pro-project bias appears to be in tension with the argument (discussed in the previous section) for placing project theories of change at the heart of impact evaluation to facilitate formulation of clear and testable impact hypotheses (cf. Ton, 2012). However, the piloting of the QUIP demonstrated that this apparent tension can at least partly be resolved by separating the role of data collection from that of analysis. In other words, an exploratory data collection stage of the QUIP was nested within, but contractually separated from a confirmatory analysis stage. The resulting channels of upward information flow are illustrated in Figure 1: solid arrows showing formal routes, and dotted arrows showing informal ones. Downward and horizontal dissemination of findings from the lead evaluator to agency staff and other stakeholders are also important, but omitted for simplicity.

The QUIP piloting illustrates how demarcation of roles can contribute to reducing pro-project bias and hence achieving greater credibility. Cognitive debriefing in Malawi after the pilot testing established that partial blinding of the field team and respondents was indeed feasible, aided by delegation of the lead evaluator role from the NGO to Bath University However, at the same time clear limits to the extent and sustainability of such blinding emerged. It was both practically necessary and ethically important for the field team and respondents to have a broad understanding of the reasons for the research. In the case of the

Malawi pilot the explicit rationale was to gain a better understanding of recent changes in rural livelihoods and food security in selected localities, and the main causes of these changes. In the light of information thereby generated it would have been easy for the lead researchers to confirm the identity of the specific project being evaluated. Hence while useful, partial blinding is ultimately not a substitute for researchers' skills, integrity and professionalism. In addition, respondents need some label to attach to visiting researchers, and if not associated with a specific project or NGO then they can be expected to ascribe another label which may also prompt strategic bias - e.g. pro-authority bias if respondents are perceived as representatives of government. Hence alongside the professionalism of researchers there is the issue of the reputation and perceived independence of the organisation to which they are affiliated – in the Malawi case a public university.

Figure 1. Organisational chart for nesting the QUIP into an impact evaluation.



Linking impact assessment and change monitoring. One limitation of relying on exploratory narrative accounts of impact (or what is also referred to above as self-reported impact assessment) is that it is unlikely respondents will generate

consistent or reliable estimates of the *magnitude* of impact on key variables, even if encouraged to do so. For this reason the QUIP was designed as a complement to ongoing quantitative monitoring of key impact indicators.[10] Repeat monitoring surveys can be used to estimate the magnitude of changes in food security, with the QUIP providing complementary evidence of different factors contributing to these changes. This can at least help to establish limits to the magnitude of change that might conceivably be attributed to an intervention. For example, if monitoring revealed that an indicator, Y_1, of household disposable income on average rose by 2% between baseline and a repeat survey, it would still be possible for the intervention to have an average impact of more than 2% if it was offsetting the negative impact of a change in some confounding variable, Z_1, such as rainfall. However, claims of impact in excess of observed changes QUIP data would also need to be substantiated by evidence that these confounding causal effects were indeed present.

Monitoring surveys also have a potentially important role to play in providing a sampling frame for QUIP sample selection, and the credibility of joint analysis is enhanced if the two sources of data are for a common set of households. However, this does present some practical challenges. First, the risk of respondent fatigue needs to be addressed by limiting the length of interviews and ensuring there is a sufficient interval between them. Second, timing and careful management of how the QUIP field team are introduced to respondents is needed to minimise the extent to which they are aware of the project being evaluated, and associated with it by the respondent.[11]

Timing of interviews and sample selection. Without precluding information about important causal processes over a longer period the QUIP primarily seeks evidence of drivers of change from respondents with reference to the period since the project being evaluated began, ideally coinciding with a baseline monitoring survey. As lead researchers are not aware of project activities it is the responsibility of the lead evaluator to confirm precisely when project activities take place and can be expected to have an impact, taking into account that this may not be uniform over the project area. QUIP studies can then be timed to minimise recall periods after the end of the farming cycle. A potential strength of the QUIP is that findings are separable and additive – i.e. each interview independently adds to understanding of impact pathways – and supplementary studies can be organised relatively quickly to extend understanding over time and across space. The link with monitoring is again important, the ideal being that monitoring of key indicators is sufficiently frequent and wide to inform decisions about when and where additional QUIP studies might be most useful. The Malawi piloting revealed the important role the lead evaluator has in obtaining sufficient detail about the implementation of project activities in order to inform the timing of QUIP studies. More specifically, the relative lack of explicit and implicit feedback on the second project was attributable to greater than anticipated variation in the roll out of activities within the project area.

An additional and relatively neglected sample selection issue is who to interview *within* households. Agricultural projects face the risk of raising yields and combined household income without necessarily also improving the welfare of all household members: for example, due to substitution of resources from food crops to cash crops, or indirect effects on gender and age specific work allocation. One way

to address this is through multiple interviews within each household to provide greater detail of information and gender sensitivity, but at the extra cost of doubling up on interviewers, and having to invest time in reconciling potentially inconsistent data. Separate second interviews within each household can also be difficult to arrange (due to absences for work, for example), and resolving differences in answers risks creating or accentuating tensions within the household. Primarily for this last ethical reason QUIP interviews during the pilot stage were limited to one per household, starting with the primary respondent identified from project lists, but without ruling out participation of other household members. At the same time the QUIP pilots augmented household data with exploratory gender/age-specific focus groups to explore whether replicating discussions within small peer groups rather than a household setting might elicit different data.[12] For example, we hypothesised that respondents might be more likely to complain about gendered effects arising from a shift to cash cropping outside their own household and without having to refer to it specifically. Table 2 provides some evidence in support of this, with the focus groups offering a proportionately higher ratio of negative to positive statements.

Interview questions and data analysis. An important design issue for the QUIP was the purpose, balance and sequencing of generative, supplementary and closed questions. Generative (more exploratory) questions are used to stimulate open-ended discussion about drivers of change (relating to each potential impact domain in turn, with optional supplementary questions used to sustain the conversation about the topic, and closed (more confirmatory) questions used to round off the

discussion of that domain before moving onto the next. This design allowed respondents to raise unknown and unexpected issues and to volunteer statements about the project without prompting, thereby making them more credible.

Linked to this question of the mix between more open and closed questions is the issue of how to balance credibility and cost of data recording and analysis. The social science 'gold standard' to record, fully transcribe, if necessary translate and inductively analyse narrative interviews within a qualitative package, such as NVivo, was rejected as too expensive to replicate for impact evaluation of relatively small projects. As an alternative the QUIP relies mainly on note-taking by the interviewer (structured according to the interview schedule), who then type these up into a pre-formatted Excel spread sheet. The research team at Bath University then explored how the data could most effectively be coded, analysed and reported. A method of highlighting and coding blocks of text in Excel was developed, so that it could then be sorted mechanically according to the domains and types of statement set out in Table 2. Ambiguity in coding definitions was identified and removed by duplicating the task using NVivo, and through exhaustive iterations of the coding and analysis. Development of this analysis protocol considerably reduced the time required for analysis of later QUIP studies using the same domains, and by doing this work in Excel rather than specialist software such as NVivo the cost of adapting the QUIP for other kinds of project and impact domains should be lower.

Conclusions

This final section draws together the argument concerning the balance between confirmatory and exploratory approaches to impact evaluation. While illustrated with reference to the development of the QUIP as a means to assessing rural development projects the main purpose of the paper is not so much to argue for or against any particular method. Rather it argues that the confirmatory/exploratory distinction in impact assessment is neglected relative to other relevant dichotomies, including quantitative/qualitative and top-down/participatory. Confirmatory approaches understandably feature more prominently in the *impact evaluation* literature where there is a narrower focus on who needs to know what about particular interventions or evaluands, whereas exploratory approaches have a closer affinity with *impact research* of a more open-ended and naturalistic kind. The former can aspire to greater rigour by adopting a hypothetico-deductive approach based on predetermined theories of change. In contrast, exploratory approaches give greater weight to avoiding bias towards project theory in the mind of both evaluators and researchers, including being blinkered from considering unintended consequences. Cost-effectiveness may appear to be best served by being able to focus impact evaluation narrowly and precisely on specific activities. But in more complex contexts there is the danger that such information will be overwhelmed or rendered redundant by contextual factors that were not even 'on the radar screen'. If so, then seemingly broader and less focused impact research may be very cost-effective indeed.

The account of experience to date with designing the QUIP highlights how these tensions relate both to overall evaluation design and to methodological details. It also suggests there may be some scope for resolving these through a nested design that encompasses a more confirmatory overall approach with more exploratory components to provide additional information.[13] The experience to date of developing the QUIP points to three sets of factors that might influence the balance between them.

First there are epistemological issues relating to the state of knowledge and understanding of the context and systems within which a project operates. The QUIP aims to be better able to reflect uncertain and insufficiently understood impact pathways, including those arising from the poorly understood interplay between different activities and processes. For example, one 'reality check' emerging from the Malawi pilots concerned the number of other development agencies and projects operating in the same area – an issue that PADev also directly addresses. This emphasis on uncertainty arising from system complexity contrasts with more quantitative approaches to impact assessment that work best when theories of change and likely outcomes are more fully and confidently understood, making it easier to codify, quantify and design experiments in advance.

Second, there are contrasting priorities over the level of resolution or kind of knowledge most valued, including a possible trade-off between numerical estimates of the average impact on intended beneficiaries, and of variation in the nature of impact *between* them. More exploratory approaches, such as the QUIP, are less useful in generating strictly quantitative estimates of the former. On the hand, they are scalable, because more interviews can be added as necessary to add to

understanding of variation in the experiences of additional respondents. Sampling and questionnaire design can also be adjusted to focus a QUIP on exploring more specific issues and/or sub-sets of intended beneficiaries, hence offering flexibility and supporting a more adaptable and cumulative approach to learning appropriate to faster moving situations.

Third, there are more sociological issues concerning the relationship between project implementers, intended beneficiaries, and those who commission and carry out impact evaluation. The QUIP aims to generate independent evidence particularly for those without opportunities to evaluate claims to impact against their own direct experience and observation of a project, and as a check against the risk of bias arising from such familiarity. The key underlying issue here is one of trust. While the QUIP seeks to elicit intended beneficiaries' accounts of how a project has affected them, it is also institutionalises suspicion in its emphasis on the extent to which such accounts can be affected by pro-project bias. One practical limitation on the scope for using QUIP may be the availability of independent researchers without direct knowledge of the project and its staff, but with necessary field skills and knowledge of local languages.

What the example of QUIP illustrates more generally is that the task of choosing between different evaluation designs entails weighing up potential sources of error and bias that are by definition uncertain. Some sources of bias – e.g. linked to statistical sampling and selection – can be quantified to some extent, and this perhaps helps to explain why they get more attention, but if so this suggests cognitive bias towards the known over the unknown. Where do these methodological preferences from? Academic background may be one source, but is

mitigated by the tendency of disciplines to sub-divide internally to accommodate methodological differences (Abbott, 2000). A deeper connection can also be made to the different ways we think, including those explored by McGilchrist's (2010). McGilchrist's work highlights one possible influence on the mental models that mould our methodological preferences in the face of uncertainty. Making the additional jump to a sociological level of analysis, they are also be reflected in *shared* mental models not only of impact evaluation but of development, instantiated in what Eyben (2013) refers to as the "institutional artefacts" governing how development organisations are structured and managed. For example, confirmatory approaches to thinking about impact evaluation are congruent with a results-based shared mental model of doing development that is certainty seeking and inclined to be more controlling. In short, methodological preferences, including the balance between confirmatory and exploratory tendencies in impact evaluation can be viewed as a symptom of a wider tension between "planners" and "seekers" in development thinking (Easterly, 2006).

Funding and acknowledgements

Production of this paper was supported under research grant ES/J018090/1 from the Department for International Development (DFID) and the Economic and Social Research Council (ESRC). I am grateful for substantial inputs from Fiona Remnant, and from participants at the project methodology workshop held in Shrewsbury in June 2013. Peter Mvulu, Myriam Volk and the spreadsheet engineers of F1F9 have also made important contributions to design and testing of the QUIP. I am also

grateful for feedback from three anonymous referees, and from participants at seminar presentations in Bath, Birmingham, London and Oxford.

Notes

[1] Other issues include the following: (a) the need to achieve a minimum sample size as a precondition for making any inference about impact at all, (b) that results comprise typical or average effects across samples or sub-samples, hiding what may be important variation within them, (c) ethical concerns about the need for involvement of respondents who receive no direct benefit, (d) the need to pre-commit to testing the outcome of a relatively small and uniform set of treatments, (e) limited possibilities for addressing heterogeneity (Vaessen, 2011; Picciotto, 2012). Good econometricians are of course aware of these issues and appropriately cautious in drawing conclusions (e.g. Heckman and Smith, 1995; Ravallion, 2009).

[2] Stern *et al.* (2012:25) also distinguish between "theory based" and "participatory" approaches, but also identify a third "case based" approach, which can to some extent be mapped onto quadrant III in Table 1.

[3] This is not to deny a continuum in the tension between confirmatory/exploratory and deductive/inductive reasoning *within* different methods. Process tracing, for example, explicitly includes a 'process induction' stage, while general elimination methodology and contribution analysis emphasise the importance of formulating and testing rival causal explanations for observed impact. The possibility of institutional bias towards approaches that are more confirmatory nevertheless remains.

[4] Also of interest, but beyond the scope of this paper is the potential for exploratory but primarily quantitative impact evaluation, including decision tree analysis and other forms of data mining (Davies, 2013).

[5] McGilchrist (2010) suggest humans are all capable of thinking in two distinct and complementary ways. The first abstracts and simplifies, producing narrower, more precise and focused models of the world. The second is associated with open forms of attention and vigilance, alongside broader, contextualizing and holistic ways of thinking. Much of the time we employ both together, and this

confers immense potential evolutionary advantages: to think narrowly (as forensic hunter-gatherer) and broadly (as agile evader of other hunters) at the same time, for example. But that does not rule individuals from having a stronger predisposition towards one way of thinking over the other. Hence, for example, more quantitative ways of thinking about impact evaluation might fit more comfortably with the first (rational, abstract, precise, generalising, certainty seeking, depersonalised), and qualitative approaches with the second (reasonable, concrete, less certain, contextual, person rather than idea oriented, emphasising difference rather than sameness). McGilchrist and Rowson (2013:30) make clear this is a completely different - if potentially complementary - distinction to that made between "fast" and "slow" thinking explored by Kahneman (2011).

[6] Following Mackie (1965), 'necessary' in this context strictly means "INUS" or Necessary but Insufficient as a particular causal package that may be Unnecessary but is Sufficient as a cause of Y.

[7] It covers work carried out between November 2012 and May 2014 as part of the three year 'ART Project' programme of research into "assessing rural transformations". This is in turn funded under a joint call of the UK Economic and Social Research Council (ESRC) and Department for International Development (DFID) for research into "measuring development".

[8] This in turn drew upon a QUIP designed during the 1990s to meet the specific needs of microfinance organisations that also linked in-depth impact interviews with routine quantitative monitoring of 'client level' indicators (see *Imp-Act,* 2004).

[9] A draft copy of the QUIP is available at http://www.bath.ac.uk/cds/projects-activities/assessing-rural-transformations/index.html

[10] In the case of SHA, it was already committed to routine monitoring of the food security of intended beneficiary households using the individual household method (IHM) developed by the NGO Evidence for Development (EFD). This approach is based on a combination of participatory rapid rural appraisal, structured household interviewing and simulation using bespoke software. Field data is used to generate estimates of how the production, exchange and transfer entitlements (in cash and kind) of a sample of households compare with estimates of their food consumption needs based on standardised nutritional requirements and food conversion ratios. Adult equivalent entitlements for a cross section of households are then compared with a benchmark absolute poverty threshold and

can be used to simulate the heterogeneous impact of price, output, income and other shocks, as well as the impact of project interventions.

[11] To avoid these risks completely the two QUIP pilot studies were not conducted in in villages covered by IHM monitoring, but in villages matched to them as closely as possible. However, it is planned to repeat the second round of pilot QUIP studies with the same households as covered by the IHM surveys.

[12] More specifically the QUIP includes four focus group discussions per study (for younger men, younger women, older men and older women), with a minimum of three people present in each and a maximum of eight.

[13] The key issue is then arguably how time and attention is divided between the two. Mayne's outline of contribution analysis for example, can accommodate exploratory work as part of Step 5 ("seeking out additional evidence" within a six step process. This seems to me to overly downplay its potential role in generating information about cause and effect that is less prone to pro-project bias (Mayne, 2012:272).

References

Abbott, J. (2000). *Chaos of the disciplines*. Chicago and London: Abbott, A., 2000 Chaos of the disciplines. University of Chicago Press: Chicago and London

Anderson, M B, Brown, D, & Jean, I. (2012). *Time to listen: hearing people on the receiving end of international aid.* Cambridge MA: CDA Collaborative Learning Projects.

Bevan, P. (2013). *Researching social change and continuity: a complexity-informed study of twenty rural community cases in Ethiopia, 1994-2015.* Mokoro Ltd. Oxford.

Cartwright, N. (2011). A philosopher's view of the long road from RCTs to effectiveness. *The Lancet, 377*(9775), 1400-1401.

Copestake, J. (2013). Research on microfinance in India: combining impact assessment with a broader development perspective. *Oxford Development Studies, 41* (Supplement), 18.

Deaton, A. (2010). Instruments, Randomization, and Learning about Development. *Journal of Economic Literature, 48*(2), 424-455. doi: 10.2307/20778731

Dietz, T. (2012). Participatory assessment of development in Africa. In N. Pouw & I. Baud (Eds.), *Local governance and poverty in developing nations* (pp. 215-239). New York: Routledge.

Dietz, T. (2013). The PaDev Story: PaDev 2007-2013 End of Project Report. Leiden: African Studies Centre.

Dietz, T., Bymolt, R., Belemvire, A., van der Geest, K., de Groot, D., Millar, D., . . . Zaal, F. (2013). PaDev Guidebook: Participatory Assessment of Development. Leiden: University of Amsterdam, Tamale University for Development Studies, Expertise pour le Developpement du Sahel, ICCO Alliance, Prisma, Woord en Daad.

Duvendack, M, Palmer-Jones, R. , Copestake, J, Hooper, L , Loke, Y , & Rao, N. (2011). *What is the evidence of the impact of microfinance on the well-being of poor people?* London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Easterly, W. (2006). *The white man's burden: why the West's efforts to aid the rest ahve done so much ill and so little good*. New York: Penguin Press.

Eyben, R. (2013). *Uncovering the politics of 'evidence' and 'results'. A framing paper for development practictioners.* . Institute of Development Studies. Brighton. Retrieved from www.bigpushforward.net

Green, D, Roche, C, Eyben, R, Dercon, S, & Witty, C. (2013). The political implications of evidence-based approaches to development. Retrieved from www.oxfamblogs.org/fp2p/?p=13344

Gulrajani, N. (2010). New vistas for development management: examining radical-reformist possibilities and potential. *Public administration and development, 30*(2), 136-148.

Haidt, J. (2012) The righteous mind: why good people are divided by politics and religion. London: Penguin.

Hammersley, M. (2013). *The myth of research-based policy and practice*. Los Angeles: Sage.

Heckman, J. J., and Smith, J. A. (1995). Assessing the case for social experiments. *The Journal of Economic Perspectives* 9, 85-110.

Holland, J (Ed.). (2013). *Who counts? The power of participatory statistics*. Brighton, UK: Institute of Development Studies.

Kahneman, D. (2011). *Thinking, fast and slow*. London: Allen Lane.

Levins, R. (1966). The strategy of model building in population biology. *American Scientist, 54*(4), 11.

Lewis, J, & Ritchie, J. (2003). Generalising from qualitative research: a guide for social science students and researchers. In J. Ritchie & J. Lewis (Eds.), *Qualitative research practice* (pp. 263-286). London, Thousand Oaks, New Delhi: Sage.

Mayne, J (2012) Contribution analysis: coming of age? *Evaluation*, 18(3):270-280.

McGilchrist, I. (2010). *The master and his emissary: the divided brain and the making of the Western World.* New Haven: Yale University Press.

Patton, M. Q. (2011). *Developmental evaluation: applying complexity concepts to enhance innovation and use*. New York: Guilford Press.

Pawson, R., & Manzano-Santaella, A. (2012). A realist diagnostic workshop. *Evaluation, 18*(2), 176-191.

Pawson, R., & Tilley, N. (1994). What works in evaluation research? *British Journal of Criminology, 34*(3), 15.

Picciotto, R. (2012). Experimentalism and development Evaluation: Will the bubble burst? *Evaluation, 18*(2), 213-229.

Ravallion, M. (2009). Should the Randomistas Rule? *The Economists' Voice* **6**, 6.

Scheifer, U. (2008). Integrated evaluation of change. Lisbon: Lisbon University Institute.

Shaffer, P. (2013). Ten Years of "Q-Squared": Implications for Understanding and Explaining Poverty. *World Development, 45*, 269-285.

Stern, E, Stame, N, Mayne, J, Forss, K, Davies, R, & Befani, B. (2012). Broadening the range of designs and methods for impact evaluations (pp. 91): DFID.

Ton, G. (2012). The mixing of methods: a three-step process for improving the rigour in impact evaluations. *Evaluation, 18*(1), 20.

van der Gaag, J, W, Gunning J, & Rongen, G. (2013). MFS II Joint evaluations synthesis report on the base line country studies. Amsterdam: Amsterdam Institute for International Development.

White, H. (2010). A contribution to current debates in impact evaluation. *Evaluation 16*(2), 11.

White, H., & Phillips, D. (2012). A*ddressing attribution of cause and effect in 'small n' impact evaluations: towards an integrated framework*. London: International Initiative for Impact Evaluation.

Woolcock, M. (2009). Towards a plurality of methods in project evaluation: a contextualised approach to understanding impact trajectories and efficacy *BWPI Working Papers*. University of Manchester: Brooks World Poverty Institute.